Originally published in: Revue Roumaine de Linguistique vol. 64 (2019) no. 3, pp. 225-226.

INTRODUCTION

RUXANDRA COSMA[1], MARC KUPIETZ[2]

The present volume is a chronicle of the making of CoRoLa, the Reference corpus of contemporary Romanian, and of its accompanying project DRuKoLA. It re-creates the timeline, describes parts of corpus architecture, components, workflow, analysis tools, the analysis platform KorAP and harmonization procedures, it spotlights aspects of its development, discusses accomplishments and problems, sketches future actions. We have comprised these steps in this volume not only to explain the complexity of the process, the outcome or interrogation ways but to connect and create links. This happens in so many different ways in this combined project, by linking what in the beginning was focused on language engineering research aims with research interests and needs of professional linguists, by linking across languages Reference Corpora of German and Romanian under one analysis platform, enabling a.o. the creation of virtual comparable corpora and contrastive analysis.

CoRoLa has two major components, a collection of annotated texts, on the one hand, and analysis tools, three of them, on the other[3].

CoRoLa is the result of conjoined forces: two research institutes of computer sciences of the Romanian Academy – in Bucharest (ICIA/RACAI, the *Mihai Drăgănescu* Research Institute for Artificial Intelligence) and Iaşi (IIT/ICS, Institute of Computer Science) – have built the corpus and developed interrogation tools for it. The central analysis tool for using CoRoLa is, however, part of a project of the Leibniz-Institute for the German Language (IDS) in Mannheim, KorAP, an analysis-platform for big language data, soon to replace the analysis instrument of DeReKo, the German reference corpus (Deutsches Referenzkorpus). The technological transfer of KorAP was possible in a different project, DRuKoLA, working in conjoined teams of the three above mentioned institutes, together with linguists from the University of Bucharest. DruKoLA focused on harmonizing Romanian data and procedures with the German analysis instrument. While CoRoLa itself is the output of a project of the Romanian Academy, which has seen it as a national research priority to make Romanian accessible to linguists and other interested users all over the world, the process of harmonizing the analysis tool to Romanian, to enable the creation of virtual comparable corpora, has been supported by a research group linkage programme of the Alexander von Humboldt – Foundation. The design behind this combined effort is to create a network of Reference Corpora of European Languages, EuReCo, that can, among other things, be used for contrastive studies. The natural language

[1] University of Bucharest, ruxandra.cosma@lls.unibuc.ro
[2] Leibniz-Institute for the German Language, Mannheim, kupietz@ids-mannheim.de
[3] The Reference Corpus of Contemporary Romanian, http://corola.racai.ro

processing of Romanian and this German-Romanian project were road-openers in EuReCo (see Kupietz *et al*., in this volume).

This volume is mainly structured into three parts, recreating the timeline of the project: 1. the creation of the Reference Corpus of Contemporary Romanian, CoRoLa, 2. the German-Romanian cooperation project DRuKoLA for the integration of the respective corpora and analysis tools, and some chalk talk on using CoRoLa, 3. linguistic applications.

The first part starts with an overview of the history, properties and usage scenarios of CoRoLa by Dan Tufiş, Verginica Barbu Mititelu, Elena Irimia, Vasile Păiş, Radu Ion, Nils Diewald, Maria Mitrofan, and Mihaela Onofrei. This overview is followed by an article that goes more into the details of design principles, text processing and cleaning, and the automatic annotations, outlined by members of the Iaşi team, in cooperation with members of the Bucharest team of CoRoLa, Daniela Gîfu, Alex Moruz, Cecilia Bolea, Anca Bibiri, and Maria Mitrofan.

At the beginning of the second part, Marc Kupietz, Ruxandra Cosma and Andreas Witt describe the project DRuKoLA. They report on its underlying ideas, history, goals, and results and argue for the necessity of close integration of research infrastructure, research data, research tools, quantitative analysis and qualitative interpretation for a productive development of empirically grounded language research. In the following paper Nils Diewald, Verginica Barbu Mititelu, and Marc Kupietz introduce the user interface of the corpus analysis platform KorAP, describing in detail its underlying design principles. In addition, they report on first experiences with Romanian linguists who have used KorAP for research in CoRoLa. In the final paper of the middle part Dan Cristea, Nils Diewald, Gabriela Haja, Cătălina Mărănduc, Verginica Barbu Mititelu, and Mihaela Onofrei explain in detail how to use one of KorAP's query languages to perform searches in CoRoLa's primary text data and its linguistic annotations. They give various usage examples covering a wide range of application scenarios and types of users.

The last two papers in this volume deal with linguistic applications. Grammatical analysis and corpus evidence can be combined in different ways and be differently balanced. The aims of these two projects, besides linguistic research targets, included using the corpus and testing queries of KorAP, observing the annotation and signaling their observations at different moments in the making of CoRoLa. Maura Cotfas discusses subjunctives and infinitive clauses in Romanian, while Cornilescu and Cosma verify predicted models of adjectival linearization.

We thank each person involved in these two projects and everyone who has used CoRoLa during its making. Most of all, we thank the institutions supporting these projects, the Romanian Academy and the Alexander von Humboldt-Foundation, we thank our text providers, listed on the website of the CoRoLa, and the hearty volunteers who have helped in the making.

It is clear to us that not everything in this volume will be easy for every reader to understand. Hopefully to a greater extent, this is because the project described here was indeed very complex and multi-directed and its success depended on bringing together experts from very different fields. Nevertheless, we apologize for the excessive use of acronyms, for which, of course, we do not blame the computer scientists among us, and hope that despite all obstacles every reader can find so much of interest in this volume that it was worth the effort in the end.